Non-separable mode dependent transforms for intra coding in HEVC

Adrià Arrufat [#*1], Pierrick Philippe [#2], Oliver Déforges [*3]

[#] *Orange Labs,* [*] *IETR/INSA*

[#] *4, Rue du Clos Courtel, 35512 Cesson-Sévigné, FRANCE,* [*] *UMR CNRS 6164, 35043 Rennes, FRANCE*

[1] adria.arrufat@orange.com
[2] pierrick.philippe@orange.com
[3] olivier.deforges@.insa-rennes.fr

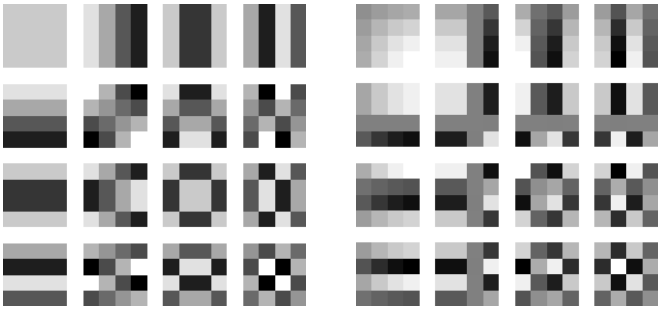
***Abstract*—Transform coding plays a crucial role in video coders. Recently, additional transforms based on the DST and the DCT have been included in the latest video coding standard, HEVC. Those transforms were introduced after a thoroughly analysis of the video signal properties. In this paper, we design additional transforms by using an alternative learning approach. The appropriateness of the design over the classical KLT learning is also shown. Subsequently, the additional designed transforms are applied to the latest HEVC scheme. Results show that coding performance is improved compared to the standard. Additional results show that the coding performance can be significantly further improved by using non-separable transforms. Bit-rate reductions in the range of 2% over HEVC are achieved with those proposed transforms.**

***Index Terms*—MDDT, KLT, transform coding, rate-distortion optimisation, adapted transform design**

I. Introduction

An important part of the design of state-of-the-art video coding standards is block-based transform coding. The Karhunen-Loève transform (KLT) is the optimal transform in terms of data decorrelation under the hypothesis of high resolution quantisation [1]. For natural images, which can be modelled as first order autoregressive Markov processes [1], the discrete cosine transform (DCT) provides a good approximation of the KLT in terms of energy compaction and performance. For this reason, and since the DCT benefits from fast algorithms, it is widely used in image and video coding standards. However, since the introduction of spatial (intra) prediction in the H.264/AVC standard, the optimality of the DCT for intra prediction residuals has been questioned.

Spatial prediction, or intra prediction (IP), makes use of different predictors to estimate the current block to encode, based on some neighbouring decoded pixels. The residual difference between the current block and the prediction is transformed and coded after quantisation. In order to improve intra coding, the mode dependent directional transform (MDDT) was introduced. The underlying idea behind the MDDT is to design an adapted transform to each prediction mode.



(a) DCT 4×4 (b) DST 4×4

Figure 1: HEVC transform bases for 4×4 TUs

The MDDT was used in [2] and [3] to improve transform coding of IP residuals. These transforms were motivated by the fact the DCT no longer approximates the KLT for this kind of blocks and a specialised transform for each IP mode was needed.

After numerous studies, an analysis in [4] reveals that the optimality of the KLT performances can nearly be achieved by using a single discrete sine transform (DST) for all IP modes. As a consequence, the DST is used in the High Efficiency Video Coding (HEVC) standard for 4×4 IP luma residuals, providing bit rate reductions of up to 1% with regards to the DCT [5].

The 2D bases for the standard transforms used in HEVC are displayed in figure 1. The first base of the DST reflects the average IP residual in HEVC, as the upper and left borders are available in intra prediction, with prediction errors increasing as they move away from the boundaries.

In this publication, the use of the KLT-based MDDT in video coding is analysed, as well as a specifically designed rate-distortion optimised transform (RDOTs), also named sparse orthogonal transforms (SOTs) in the literature. Transform separability is also questioned in this paper. Designing and testing non-separable transforms will allow finding out

the performance impact due to separability under different transform design criteria.

II. Rate-distortion optimised transform

The need of RDOTs was introduced in [6] and explained more in detail in [7].

The design of the RDOT differs from that of the KLT in the fact that RDOT no longer assumes a high quantisation resolution, where all transformed coefficients are presumably transmitted. Therefore, the KLT design does not fit the behaviour of current video coders, which implies lots of coefficients with zero values. As a consequence, the RDOT design features a bit rate constraint to increase sparsity in the transform domain.

The proposed method is able to find the optimal transform in an iterative fashion for some learning residuals data, with an initial transform and a constraint on the coefficients sparsity, as explained below.

$$\mathbf{A}_{opt} = \arg\min_{\mathbf{A}} \sum_{\forall i} \min_{\mathbf{c}_i} \left(\|\mathbf{x}_i - \mathbf{A}^T \mathbf{c}_i\|_2^2 + \lambda \|\mathbf{c}_i\|_0 \right) \quad (1)$$

Where $\mathbf{x}_i$ is a block of the training set, $\mathbf{c}_i$ are its quantised transformed coefficients using the transform $\mathbf{A}$. $\mathbf{A}^T$ is its transposed matrix, as $\mathbf{A}$ is chosen orthonormal. The constraint in the cost function is the ℓ_0 norm of the coefficients, i.e. the number of non-zero coefficients. The Lagrange multiplier λ of the constraint only depends on the quantisation accuracy applied to the coefficients, as demonstrated in [6].

The suggested design involves an iterative algorithm where the optimal coefficients are found for a given transform. Then, the transform is updated to match the optimal coefficients. Those two steps are performed until convergence, when the value of the metric is stabilised.

Transform design from equation (1) outputs non-separable transforms. Due to high requirements in terms of memory and computational power, separable transforms are exclusively used in image and video coding schemes.

For a given block $\mathbf{x}$, the transformed coefficients $\mathbf{X}$ using separable transforms are defined as:

$$\mathbf{X} = \mathbf{A}_v \left(\mathbf{A}_h \mathbf{x}^T \right)^T = \mathbf{A}_v \mathbf{x} \mathbf{A}_h^T \quad (2)$$

Where $\mathbf{A}_h$ is the horizontal transform, applied to the rows of $\mathbf{x}$, and $\mathbf{A}_v$ is the vertical transform, applied to the resulting columns.

Using the definition of the separable transforms from equation (2), the RDOT design from equation (1) was updated to compute separable RDOTs:

$$\underset{opt}{\mathbf{A}_v, \mathbf{A}_h} = \arg\min_{\mathbf{A}_v, A_h} \sum_{\forall i} \min_{\mathbf{c}_i} \left(\|\mathbf{x}_i - \mathbf{A}_v^T \mathbf{c}_i \mathbf{A}_h\|_2^2 + \lambda \|\mathbf{c}_i\|_0 \right) \quad (3)$$

Where $\mathbf{A}_v$ and $\mathbf{A}_h$ are the vertical and horizontal transforms, respectively.

The appropriateness of equation (3) for video coding has been verified in [7], where the MDDT (based on separable KLTs) was significantly outperformed by separable RDOTs, called mode dependent sparse transform (MDST).

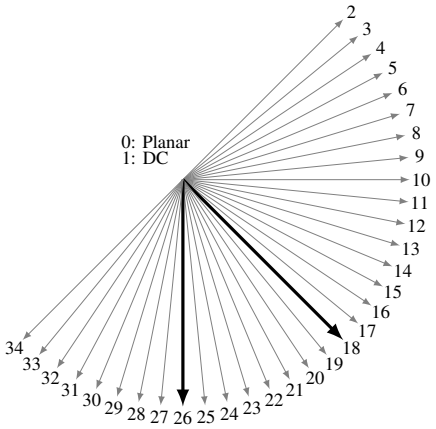
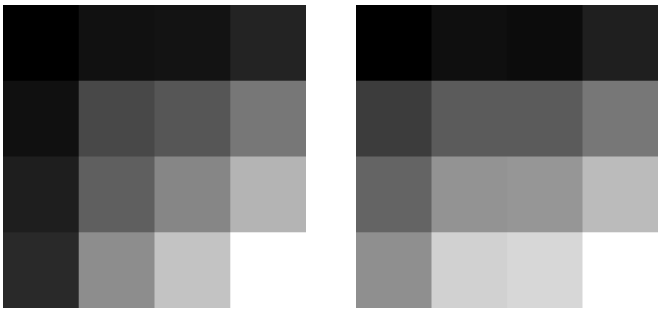


Figure 2: The 35 IP modes in HEVC



(a) IP 18 (b) IP 26

Figure 3: Average 4×4 IP residuals for highlighted IP modes

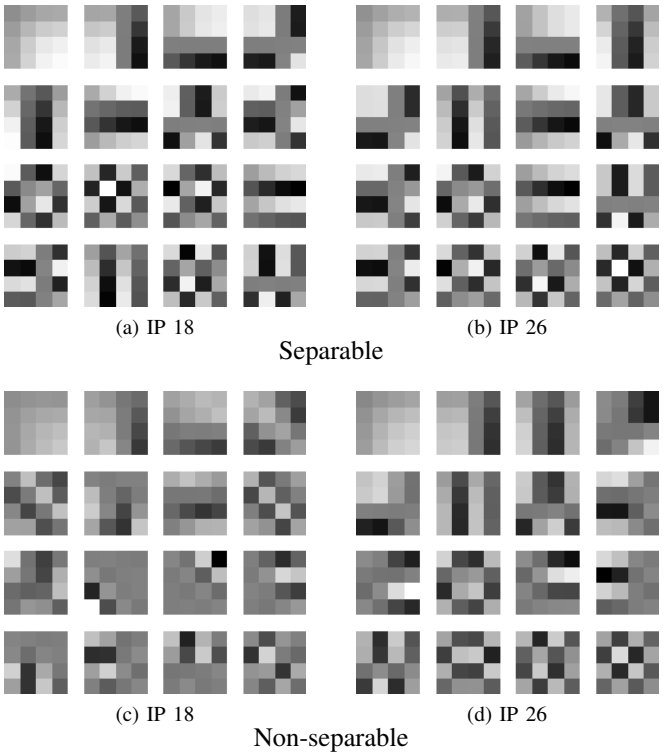
III. Non-separable transforms

Due to complexity reasons, non-separable transforms have systematically been discarded in favour of their separable counterparts. However, non-separable transforms might be better at compacting the signal on fewer transform coefficients, since they can exploit any individual correlation between pixels within a block.

For illustrative purposes, figure 3 displays the average residual magnitudes for IP directional mode 18 and 26 as selected by HEVC. The prediction directions are highlighted in figure 2. Dark colours in figure 3 represent low values, hence good predictions, whereas light colours indicate where important errors in prediction tend to occur. When IP mode 18 is selected, good predictions are made along the top and the left borders of the block. If the IP mode 26 is used, predictions are based only from the upper boundary. In both cases, errors increase with the distance to the top-left boundaries.

Figure 4 displays the obtained bases for IP modes 18 and 26, sorted by decreasing average coefficient magnitude. While mode 26 is purely vertical, mode 18 has a strong diagonal component, which can be seen by how the bases adapt to the nature of the IP residuals.

The first separable RDOT bases for all IP modes tend to be similar to those of the DST-VII used in HEVC, displayed in figure 1-b, as well as those coming from non-separable RDOTs



(a) IP 18 (b) IP 26

Separable

(c) IP 18 (d) IP 26

Non-separable

Figure 4: 4×4 RDOTs for highlighted IP modes

whose predictions are either horizontal or vertical. This is a reassuring fact, as the DST-VII has been proved to be the nearly optimal choice if only one transform is used for all prediction modes.

It is important to notice that the prediction direction can be spotted in some of the bases in figure 4-c. That kind of patterns are not achieved using separable transforms.

This fact has motivated the following experiments: a non-separable MDDT based on the RDOT from equation (1) to unveil the performance gap caused by separability.

IV. Experimental Results

In order to find out the gains obtained by RDOTs and non-separability, the following systems have been implemented and tested:

(a) a non-separable KLT-based MDDT, named NS-MDDT
(b) a separable RDOT-based MDDT, named S-MDST
(c) a non-separable RDOT-based MDDT, named NS-MDST

All systems use one adapted transform in each IP mode. The only difference is in the transform learning method between (a) and (c) and the separability between (b) and (c).

Figure 5 illustrates how a block is decoded using the mode dependent transform scheme. The IP mode is used to generate the prediction of the current block and to select the appropriate transform. This transform is applied on the coded residual to be added to the predicted signal and reconstruct the image pixels.

To determine the optimal transforms, residuals coming from different sequences, with different resolutions and quantisation parametres (QPs) have been used to obtain one separable and one non-separable RDOT per IP mode. Afterwards, these transforms have been used in the HEVC reference software (HM version 10.1) to replace the DST for the 4×4 IP luma residuals and the DCT in the 8×8 case.

	Class	TU size: 4×4 S-MDST	TU size: 4×4 NS-MDST	TU size: 8×8 S-MDST	TU size: 8×8 NS-MDST
A	(2560×1600)	-1.41%	-1.60%	-1.55%	-2.09%
B	(1920×1080)	-0.10%	-0.52%	-0.41%	-1.66%
C	(832×480)	-0.27%	-1.41%	-0.35%	-2.67%
D	(416×240)	-0.23%	-1.18%	-0.20%	-1.38%
	Average	**-0.50%**	**-1.18%**	**-0.63%**	**-1.95%**

Table I: Average separability impact on different TU sizes

In order to adapt to HEVC's mode dependent coefficient scanning and context adapted binary arithmetic coding [5], transform bases have been re-ordered accordingly, so that HEVC scans coefficients in an increasing average magnitude order.

Systems have been tested by coding the HEVC sequence set at QPs 22, 27, 32 and 37, as specified in this standard's common test conditions [8].

The first experiment was focused on finding out the separability impact on the learnt transforms. S-MDST and NS-MDST have been compared against standard HEVC for 4×4 and 8×8 transform unit (TU) sizes. Table I contains the average Bjøntegaard distortion-rate (BD-rate) savings per class.

This table shows that, compared to HEVC, even using specifically designed RDOTs, separable transforms show modest BD-rate savings with regards to HEVC. However, performances improve substantially across all classes by using non-separable RDOTs.

The second experiment aims at comparing a complete system using mode dependent RDOTs for both TU sizes of 4×4 and 8×8 against a non-separable KLT-based MDDT. The detailed results are presented in table II.

The numbers confirm the non-separable MDST outperforms the separable MDST by achieving BD-rate savings of over 2%, while the separable MDST savings are around 0.5%.

Comparing the NS-MDDT to the NS-MDST extends the results in [7] for non-separable transforms, as the NS-MDST exhibits higher bit-rate savings than the NS-MDDT: this confirms the appropriateness of the transform design method.

The interest of non-separable transforms can be seen while analysing the detailed results. Sequences with strongly diagonal patterns, like *BasketballDrill*, are able to get BD-rate improvements of up to 8.20%.

Despite designing the transforms using only IP residuals, the system has also been tested in a random access (RA) configuration, where it outperforms HEVC by 1.73%. Improvements are possible in this configuration since the blocks coded in inter mode also take advantage of the increased quality of the intra predicted blocks.

Regarding the complexity of the proposed system, a factor of two can be observed in the encoding time, compared to HEVC. The decoding time has been increased by 30%.

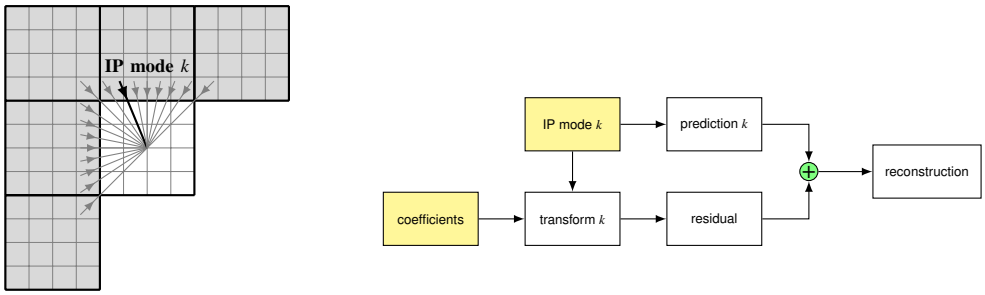


Figure 5: Decoding scheme using mode dependent transforms

	Sequence	Y BD-rate (AI) NS-MDDT	S-MDST	NS-MDST	Y BD-rate (RA) NS-MDST
Class A (2560×1600)	PeopleOnStreet	-2.59%	-0.78%	-2.09%	-0.90%
	Traffic	-2.59%	-0.91%	-2.22%	-8.84%
	NebutaFestival	0.37%	0.37%	0.37%	-0.47%
	SteamLocomotiveTrain	-3.50%	-3.50%	-3.50%	-1.54%
	Average	**-2.05%**	**-1.20%**	**-1.86%**	**-2.94%**
Class B (1920×1080)	BasketballDrive	-0.75%	-0.20%	-1.87%	-1.22%
	BQTerrace	-2.45%	-0.87%	-3.64%	-1.83%
	Cactus	-1.84%	-0.65%	-2.18%	-1.41%
	Kimono1	-0.86%	-0.27%	-0.81%	-0.31%
	ParkScene	-1.54%	-0.55%	-1.16%	-0.70%
	Average	**-1.49%**	**-0.51%**	**-1.93%**	**-1.09%**
Class C (832×480)	BasketballDrill	-5.18%	-0.93%	-8.20%	-4.47%
	BQMall	-0.56%	-0.31%	-1.70%	-0.99%
	PartyScene	-0.70%	-0.39%	-1.51%	-1.06%
	RaceHorses	-2.06%	-0.69%	-2.47%	-1.31%
	Average	**-2.12%**	**-0.58%**	**-3.47%**	**-1.96%**
Class D (416×240)	BasketballPass	-0.93%	-0.40%	-2.19%	-1.09%
	BQSquare	-0.56%	-0.26%	-1.64%	-0.89%
	BlowingBubbles	-1.12%	-0.16%	-2.05%	-1.10%
	RaceHorses	-2.64%	-0.77%	-3.09%	-1.25%
	Average	**-1.31%**	**-0.40%**	**-2.24%**	**-1.08%**
All classes	**Overall**	**-1.74%**	**-0.67%**	**-2.38%**	**-1.73%**

Table II: Comparison of separable and non-separable mode dependent transforms

However, transforms have been implemented as direct integer matrix multiplications, without any optimisation, for this reason it is believed that complexity can be notably reduced. Current complexity values can be seen as an upper bound of a final optimised system.

V. Conclusion

This paper has corroborated the adequateness of the RDOT over the KLT transform design method for video coding, as reported in [7].

Questioning the need of separable transforms and designing non-separable transforms has allowed unveiling an important performance gap. Non-separable transforms are particularly useful for sequences with highly diagonal patterns, where separable transforms cannot reach this level of performance due to their reduced energy compaction for the diagonal patterns.

The experiments run in this work provide encouraging results to extend the current system, which only uses TU sizes of 4×4 and 8×8, to use larger TU sizes, such as 16×16 and 32×32.

Even though the complexity of the encoder and the decoder have been increased, further investigations are in the works to simplify the system while having minimal impact to the currently obtained level of performance.

References

[1] E. K. R. Rao and P. Yip, *The Transform and Data Compression Handbook*. Boca Raton, 2001.

[2] Y. Ye and M. Karczewicz, "Improved intra coding," ITU-T Q.6/SG11, VCEG, Shenzhen, China, Tech. Rep. VCEG-AG11, October 2007.

[3] ——, "Improved H.264 intra coding based on bi-directional intra prediction, directional transform, and adaptive coefficient scanning," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, 2008, pp. 2116–2119.

[4] J. Han, A. Saxena, and K. Rose, "Towards jointly optimal spatial prediction and adaptive transform in video/image coding," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 726–729.

[5] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012.

[6] O. Sezer, O. Harmanci, and O. Guleryuz, "Sparse orthonormal transforms for image compression," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, 2008, pp. 149–152.

[7] O. G. Sezer, "Data-driven transform optimization for next generation multimedia applications," Ph.D. dissertation, Georgia Institute of Technology, 2011.

[8] F. Bossen, "Common test conditions and software reference configurations," ITU-T, Geneva, Switzerland, Tech. Rep. JCTVC-I1100, May 2012.